УДК 519.767:004.93

МЕТОД ЛАТЕНТНО-СЕМАНТИЧНОГО АНАЛІЗУ ДЛЯ ВИЗНАЧЕННЯ МОТИВАЦІЙ СТУДЕНТІВ В ПАРАДИГМІ СТУДЕНТОЦЕНТРОВАНОГО НАВЧАННЯ

Дворник В. А.

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Україна, Київ

Розглянута задача перевірки мотивації студентів на основі їх мотиваційних листів, визначення галузі знань та переліку дисциплін для побудови індивідуального навчального плану студента. Побудована математична модель задачі перевірки мотивацій студента. Описаний метод латентно-семантичного аналізу тексту, який представлений мотиваційним листом студента. Розроблений алгоритм та поданий опис діаграми компонентів програмної системи

Ключові слова: індивідуальна освітня траєкторія, мотивація, компетентність, область знань.

Дворник В. А. Метод латентно-семантического анализа для определения мотиваций студентов в процессе студентоцентрированного обучения / Национальный технический университет Украины "Киевский политехнический институт имени Игоря Сикорского", Украина, Киев

Рассмотрена задача проверки мотивации студентов на основе их мотивационных писем, определения области знаний и перечня дисциплин для построения индивидуального учебного плана студента. Построена математическая модель задачи проверки мотиваций студента. Описан метод латентно-семантического анализа текста, который представлен

мотивационным письмом студента. Разработан алгоритм и дано описание диаграммы компонентов программной системы.

Ключевые слова: индивидуальная образовательная траектория, мотивация, компетентность, область знаний.

V. A. Dvornyk Method of latent-semantic analysis for determining the motivations of students in the process of student-centered education / National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, Kiev

The task of checking the motivation of students on the basis of their motivational letters, the definition of the field of knowledge and the list of disciplines for the construction of the individual curriculum of the student is considered. A mathematical model of the problem of checking student motivations is constructed. The method of latent-semantic analysis of the text, which is presented by the student's motivational letter, is described. The algorithm is developed and the description of the components diagram of the software system is given.

Keywords: individual educational trajectory, motivation, competence, area of knowledge.

Вступ. На сьогоднішній день все популярнішою стає модель індивідуального підходу в навчанні, який спрямований на підтримку ефективності самого процесу навчання. Індивідуалізація навчання відбувається шляхом реалізації індивідуальних освітніх траєкторій і складає сутність студентоцентрованого навчання [1]. Реалізація студентоцентрованого навчання і розробка індивідуальних освітніх траєкторій навчання з урахуванням мотивацій, психологічних, професійних та особистісних якостей студента є сучасним трендом розвитку освіти.

Мета статті. Метою цієї роботи є обґрунтування доцільності перевірки мотивацій студентів для побудови індивідуальних освітніх траєкторій на основі їх мотиваційних листів, визначення областей знань, на опанування яких вмотивовані студенти, визначення ключових слів, які формулюють склад областей знань та побудови списку дисциплін, які формуватимуть необхідні студентові компетентності в індивідуальному навчальному плані.

Для досягнення цілей застосовуються методи автоматичної обробки текстів природною мовою, якими є мотиваційні листи студентів, психологічні, професійні та особистісні якості студентів, отримані в результаті їх тестувань. З урахуванням ролі інформаційних технологій у розвитку суспільства знань та економіки держави розглядатимемо модель освітнього процесу для галузі знань інформаційні технології.

Постановка задачі. Нехай задана колекція з n документів, набір яких представляє собою мотиваційні листи претендентів, і m різних термінів, видобутих із словника ІТ-термінів. Під терміном розуміється окреме слово або словосполучення. Для застосування моделі латентно-семантичного аналізу потрібно побудувати $m \times n$ матрицю X «термін-документ», значення x_{ij} якої містять вагові коефіцієнти терміна $t_i, i = \overline{1,m}$ в документі $d_j, j = \overline{1,n}$. Стовпці матриці X відповідають мультимножині слів для документа, при цьому можуть бути використані терміни, «зважені» за якою-небудь мірою, наприклад, TF-IDF або такі, що засновані на ентропії. Отримана матриця X «термін-документ» являє собою просторово-векторну модель подання текстової інформації природною мовою і є вхідними даними для методу латентно-семантичного аналізу [2, 3]. Потрібно визначити відповідність вхідного тексту, що поданий природною мовою, тематиці ІТ-галузі та визначити тему цього тексту.

Опис методу розв'язання задачі. Після побудови матриці X «термін-документ» застосовується сингулярне розкладання (SVD) вихідної матриці X на три матриці: $X = USV^t$. В розкладанні матриці U і V є ортогональними матрицями вимірності $m \times r$ і $n \times r$ відповідно, а матриця S – це $r \times r$ діагональна матриця, яка містить власні значення. Кожне з r власних значень матриці S відповідає одному з компонентів, що відслідковуються в колекції документів, і позначає, наскільки цей компонент актуальний у всій колекції.

Власні значення упорядковано відповідно до діагоналі матриці S в порядку спадання, так що ті власні значення, які йдуть першими, пов'язані з найбільш важливими компонентами. Це дозволяє зменшити найменш важливі компоненти до числа $k \le r$, просто видаливши відповідні рядки і стовпці в матрицях. Таке скорочення потенційно дозволяє видалити «шум» в даних, який може бути складений, наприклад, з термінів або груп, що з'являються тільки в декількох документах і погано пов'язані з іншими.

Як тільки таке значення k встановлено, можна розрахувати матрицю X' шляхом перемноження трьох усічених матриць: результуюча матриця X' матиме свій ранг, зменшений від r до k. Матриця X' структурно ідентична матриці X (її рядки і стовпці є уявленнями для тих самих термінів і документів, що і в матриці X), але вагові коефіцієнти скориговані так, що «шум» усунутий, і враховано очевидні взаємозв'язки між термінами (або між документами). Наприклад, якщо два терміни t_a і t_b часто зустрічаються разом в документах, то документ, який містить тільки термін t_a з цих двох термінів, буде в будь-якому випадку мати вагу для терміна t_b більше нуля (і навпаки).

З відновленої матриці X' або безпосередньо з усічених матриць, використовуваних для її обчислення, схожість між термінами

і між документами може бути обчислена відповідно до скоригованих вагових коефіцієнтів, які в загальному випадку будуть відрізнятися від відповідних ваг, обчислених з вихідної матриці. У загальному випадку, коли знайдені документи, які найбільше задовольняють запиту, використовується підхід, в якому запит задається як документ, який зіставляється з якимось відомим документом. Він повинен бути спочатку відображений в прихований (латентний) простір для того, щоб пройти таку саму корекцію значень (ця процедура відома як згортка). У прихованому просторі можна знайти пов'язані документи, які не містять точних слів запиту, але при цьому строго відповідають їм [2].

Для реалізації алгоритму класифікації в алгоритмі семантичного аналізу найбільш ефективною виявилась міра TF-IDF. Тут TF – частота слова, яка розраховується як відношення кількості входжень деякого терміну до загальної кількості слів документа. Таким чином оцінюється важливість терміна $t_i, i=\overline{1,n}$ в межах окремого $d_j, j=\overline{1,m}$ документа: $tf(t,d)=n_t/\sum_k n_k$, де n_k – кількість входжень терміна t в документ d, а у знаменнику – загальна кількість слів у даному документі IDF – зворотна частота документа, тобто інверсія частоти, з якою окреме слово зустрічається в документах загальної колекції. Обчислення зворотної частоти документа зменшує вагу слів широкого вживання. Для кожного унікального слова в межах конкретного набору документів існує тільки лише значення IDF: $idf(t,D)=\log(|D|/|\{d_i \in D \mid t \in d_i\}|$, де $|D|$ – кількість документів в колекції, $|\{d_i \in D \mid t \in d_i\}|$ – кількість документів із колекції D, в яких зустрічається термін t (коли $n_t \neq 0$).

Вибір основи логарифма в формулі не має значення, оскільки зміна основи призводить до зміни ваги кожного конкретного слова на сталий множник, що не впливає на співвідношення ваг.

Отже, міра TF-IDF – це добуток двох множників:

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D)$$

Звідси слідує, що велику вагу в TF-IDF отримуватимуть слова з високою частотою в межах визначеного документа і з низькою частотою використань в інших документах [2].

Опис алгоритму. Схема алгоритму латентно-семантичного аналізу [3] складається з таких кроків.

Крок 0. З аналізованих документів виключити стоп-слова, які не несуть в собі смислового навантаження: сполучники, частки, прийменники і безліч інших слів. Перейти на крок 1.

Крок 1. В аналізованих документах відфільтрувати цифри, окремі букви і розділові символи. Перейти на крок 2.

Крок 2. Виконати операцію стемінг для отримання основи слова, застосувавши її до усіх слів документів. Перейти на крок 3.

Крок 3. Побудувати частотну матрицю, стовпці якої відповідають документам, а рядки – індексованим словам. Кожна комірка матриці визначає кількість повторів слів у відповідному документі. Перейти на крок 4.

Крок 4. Нормалізувати отриману частотну матрицю за допомогою міри TF-IDF. Перейти на крок 5.

Крок 5. Провести сингулярне розкладання отриманої матриці. Якщо використовується двовимірне сингулярне розкладання перейти на крок 6.

Крок 6. Вилучити останні стовпці матриці U і останні рядки матриці Vt, залишивши тільки перші два, які відповідають координатам x,y кожного слова для матриці U і координатам x,y для

кожного документа в матриці Vt. Розкладання такого виду називають двовимірним сингулярним розкладанням. Перейти на крок 7

Крок 7. Підготувати вихідні дані у вигляді вкладених списків координат x,y для двох стовпців матриці U слів і двох рядків матриці Vt документів. Перейти на крок 8

Крок 8. Порівняти координати слів заданого словника (у випадку задачі перевірки вмотивованості) або термами областей знань (у випадку задачі визначення дисциплін) з відомими документами за допомогою косинусної відстані. Робота алгоритму закінчена.

Реалізація алгоритму. Розглянутий алгоритм реалізований у веб-застосуванні [4] в модулі формування областей знань, на які мотивовані студенти відповідно до їх мотиваційних листів. Діаграма компонентів розробленого програмного забезпечення подана на рисунку 1. Веб-застосування дає можливість студенту пройти психологічний тест, тест на лідерство і тест на сформованість професійних компетентностей, вибрати рекомендовану програмною системою ІТ-професію, перевірити мотивації студента до опанування знань та вмінь за вибраною професією, сформувати індивідуальну траєкторію навчання, яка складається із вибраних дисциплін індивідуального навчального плану.

Висновки. В роботі розглянуто застосування методу латентно-семантичного аналізу текстів природної мови для задачі перевірки мотивацій студента в процесі формування індивідуальної освітньої траєкторії. Описана змістовна постановка задачі та побудована математична модель. На основі методу латентно-семантичного аналізу тексту розроблений алгоритм, який покладений в основу модулю формування областей знань, на які мотивовані студенти відповідно до їх мотиваційних листів.

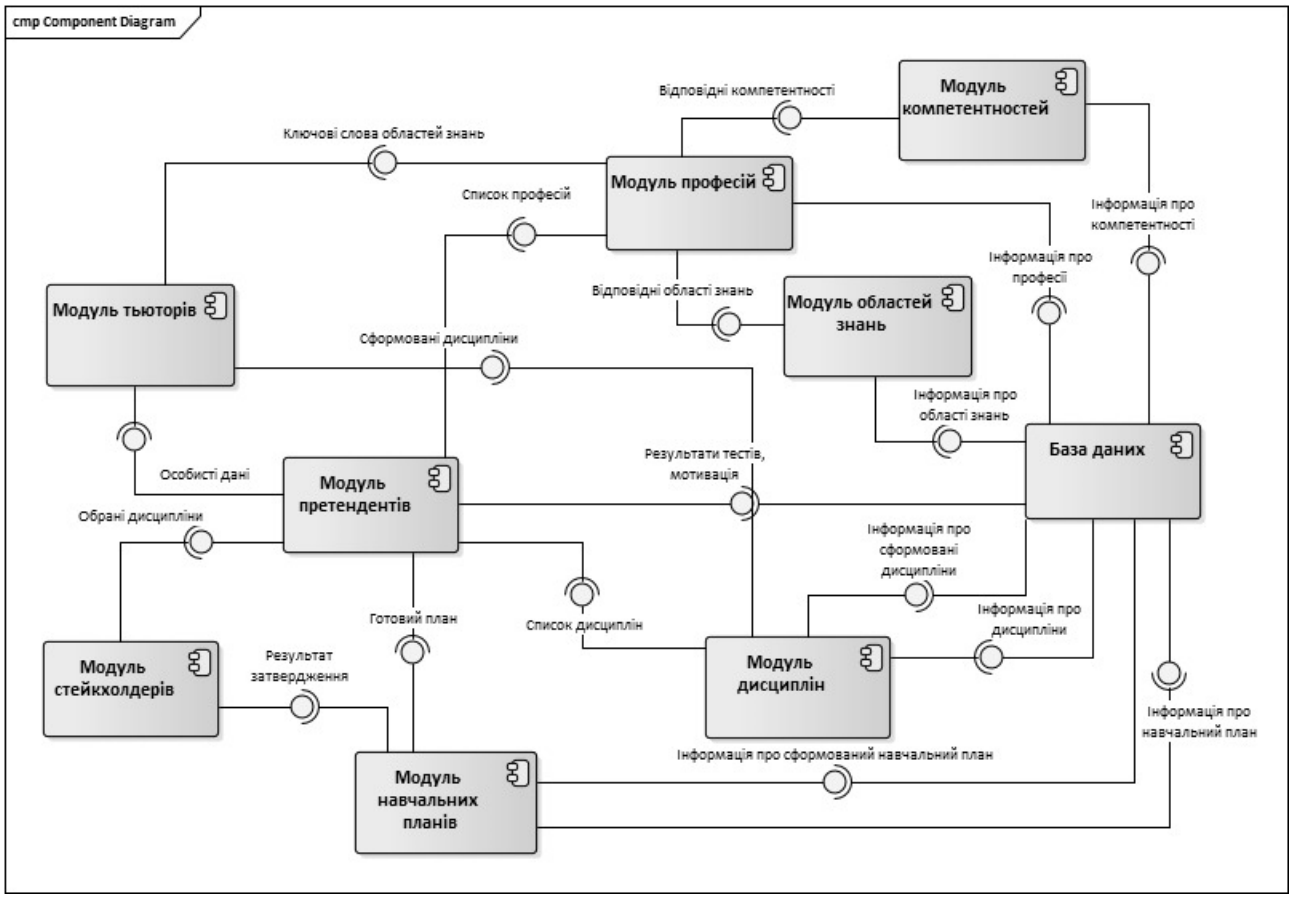

Рис. 1. Схема структурна компонентів системи побудови індивідуальної освітньої траєкторії

Література:

1. Мухаметзянова Ф. Г., Забирова Р. В. Проектирование индивидуальной образовательной траектории и маршрута студента вуза - будущего бакалавра. Казанский педагогический журнал. 2015 [Електронний ресурс], Режим доступу: https://cyberleninka.ru/article/v/proektirovanie-individualnoy-obrazovatelnoy-traektorii-i-marshruta-studenta-vuza-buduschego-bakalavra

2. Хомоненко А. Д., Краснов С. А. Применение метода латентно-семантического анализа для автоматической рубрикации документов. Известия Петербургского университета путей сообщения. № 2. 2012. [Електронний ресурс], Режим доступу:

https://cyberleninka.ru/article/n/primenenie-metoda-latentno-semanticheskogo-analiza-dlya-avtomaticheskoy-rubrikatsii-dokumentov

3. Воронин В. М., Курицин С. В. Латентный семантический анализ и понимание текста. 2010. [Електронний ресурс], Режим доступу: http://elar.urfu.ru/bitstream/10995/4085/3/pv-03-09.pdf

4. Тараненко Ю. Визуализация результатов латентно-семантического анализа средствами Python. 2017. [Електронний ресурс], Режим доступу: https://habr.com/post/335668/

References:

1. Mukhametzyanova F. G. and Zabirov R. V. Designing an individual educational trajectory and route for a university student - the future bachelor. Kazan Pedagogical Journal. 2015 [Electronic resource], Access mode: https://cyberleninka.ru/article/v/proektirovanie-individualnoy-obrazovatelnoy-traektorii-i-marshruta-studenta-vuza-buduschego-bakalavra

2. Khomonenko A. D. and Krasnov S. A. Application of the method of latent-semantic analysis for automatic document categorization. Proceedings of the Petersburg University of Communications. № 2. 2012. [Electronic resource], Access mode: https://cyberleninka.ru/article/n/primenenie-metoda-latentno-semanticheskogo-analiza-dlya-avtomaticheskoy-rubrikatsii-dokumentov

3. Voroni V. M. and Kuritsin S. V. Latent semantic analysis and text understanding. 2010. [Electronic resource], Access mode: http://elar.urfu.ru/bitstream/10995/4085/3/pv-03-09.pdf

4. Taranenko Y. Visualization of latent semantic analysis results using Python tools. 2017. [Electronic resource], Access mode: https://habr.com/post/335668/